Cell

Leading Edge

Commentary

Precision medicine in 2030—seven ways to transform healthcare

Joshua C. Denny[1,3,*] and Francis S. Collins[2]

[1]*All of Us* Research Program, National Institutes of Health, Bethesda, MD, USA
[2]National Institutes of Health, Bethesda, MD, USA
[3]Present address: Bldg. 1 Room 228, 1 Center Drive, Bethesda, MD 20814, USA
*Correspondence: joshua.denny@nih.gov

https://doi.org/10.1016/j.cell.2021.01.015

Precision medicine promises improved health by accounting for individual variability in genes, environment, and lifestyle. Precision medicine will continue to transform healthcare in the coming decade as it expands in key areas: huge cohorts, artificial intelligence (AI), routine clinical genomics, phenomics and environment, and returning value across diverse populations.

Ever since the completion of the first human genome sequence in 2003, clinicians have anticipated a data-driven transformation in healthcare. New troves of molecular and phenotypic interrogation would lead to refined diagnoses, more rational treatment, and prevention of disease. In 2011, an *ad hoc* committee at the National Research Council argued for a "new taxonomy of human diseases" based on the emerging field of precision medicine (US National Research Council, 2011).

Today, some of that promise has already been realized. Researchers are routinely using healthcare data for discovery, identifying genomic underpinnings of cancer and many other common and rare diseases, introducing transformative molecularly targeted therapies, and leveraging massive computational capabilities with new machine learning methods. We are beginning to see the fruits of these efforts.

There is perhaps no more poignant example than the response to the COVID-19 pandemic. Genomics and molecular technologies were key in identifying the etiologic agent, developing diagnostics and treatments, and creating vaccine candidates. Rapid case reporting quickly exposed vast health disparities with COVID-19 and highlighted the importance of capturing a more detailed understanding of social determinants of health. Large-scale consortia based on healthcare data quickly assembled huge datasets for rapid investigations of risk factors and outcomes, demonstrating the power of amalgamated healthcare data. Pooling data from existing research cohorts enabled rapid genomic studies that have identified loci associated with disease susceptibility and patient outcomes. COVID-19 has also called attention to the need for longitudinal cohorts to identify clinical and biologic risk factors and long-term sequelae for acute infectious disease. Many of the elements of the response to COVID-19 are basic capabilities underpinning precision medicine.

At the same time, COVID-19 has highlighted the need for precision medicine to move further and faster. In this paper, we suggest seven opportunities to accelerate an equitable realization of the promise of precision medicine (Figure 1). Their impacts are outlined in Table 1.

Huge, interoperable, longitudinal cohorts

Over the last two decades, national cohorts such as the UK Biobank, the Million Veteran Program, FinnGen, and the *All of Us* Research Program have amassed huge populations with genomic, laboratory, and lifestyle assessments as well as longitudinal follow-up on health outcomes. The depth and breadth of the data are staggering, as are the opportunities for discovery across every area of medicine.

In order to maximize the impact of these resources, an "open science" approach is emerging. For example, the UK Biobank has opened its doors to more than 19,000 "bona fide researchers" from 80 countries, and researchers can start using the *All of Us* Research Program's data cloud in as little as two hours after initial login.

The next step is clear: make it easier for researchers to merge data from multiple cohorts. Currently, this requires painstaking manual phenotype adjudication and building large consortia including experts from each cohort. Fortunately, there are efforts underway to improve this process. Groups such as the Global Alliance for Genomics and Health (GA4GH) are working to develop and to coordinate common data models and file formats to facilitate collaboration and interoperability. In recognition of the need for better collaboration, the International Hundred Thousand Plus Cohort Consortium (IHCC) has brought together more than 100 cohorts in 43 countries comprising more than 50 million participants—nearly two orders of magnitude bigger than the biggest single cohort today (Manolio et al., 2020). It would be hard to overstate the impact this work could have on global research efforts.

Improved diversity and inclusion in science

One of the biggest challenges (and opportunities) before the biomedical enterprise today is the lack of diversity in populations involved in research studies. Less than 3% of the participants in published, genome-wide association studies are of African or Hispanic or Latin American ancestries, and 86% of clinical trial participants are white (Knepper and McLeod, 2018; Mills and Rahal, 2020). The lack of diversity in research risks exacerbating

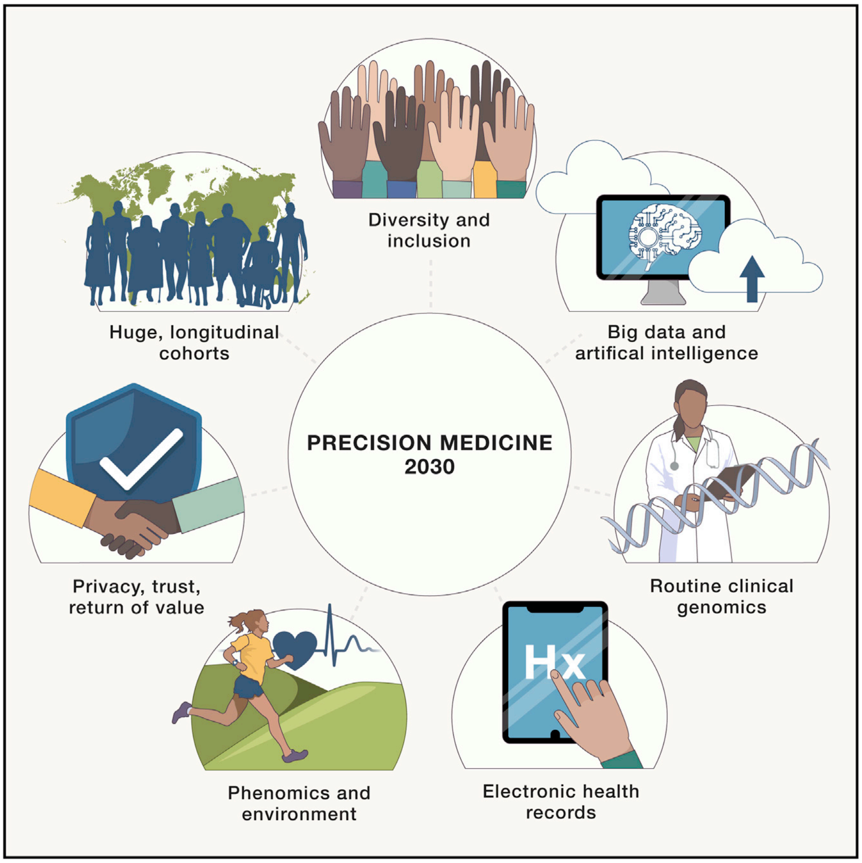
Diversity and inclusion
Huge, longitudinal cohorts
Big data and artifical intelligence
PRECISION MEDICINE 2030
Privacy, trust, return of value
Routine clinical genomics
Hx
Phenomics and environment
Electronic health records

Figure 1. Seven opportunities for precision medicine by 2030

health disparities and also impoverishes biologic discovery that could be applicable to all populations.

With a growing depth of data, we have an opportunity to replace adjustments for race and ethnicity with more specific measures. In particular, "race" conflates a plethora of social, cultural, political, geographic, and biologic factors together and can perpetuate systemic racism. Routine collection of social determinants of health in both research and clinical care in combination with more precise measures of environmental influences, habits, and genetic ancestry can provide more rational, etiology-based adjustments and yield better risk stratifications and treatments (Wilkins et al., 2020).

As we work toward increasing the diversity of populations in studies, we should also increase the diversity of the biomedical research workforce. A more diverse workforce—in culture, ancestry, beliefs, scientific backgrounds, and methodological approaches—brings increased understanding, innovation, trust, and cultural sensitivity; is more likely to pursue questions relevant to different audiences; and ultimately delivers better research (Hofstra et al., 2020).

As international collaborations grow, researchers will also need to consider the ethics of international collaboration and rotate leadership, authorship, and resources to ensure that research benefits developing countries as well as more advantaged ones. Establishing international infrastructures and science facilities—not just access to samples and data—will produce long-term benefits that accelerate health and capabilities.

Big data and artificial intelligence

Big data and artificial intelligence (AI) are transforming previously intractable problems such as search optimization, language translation, image interpretation, and autonomous driving. Many accrued biomedical data sets meet all "5 V's" of big data since they are voluminous, high velocity, come in many varieties, have significant variability, and have intrinsic value. However, AI approaches in medicine have been limited by the (un)availability of large, commonly structured datasets.

Looking forward, biomedical datasets will become increasingly ready for analyses. As we discuss in the following sections, the growth of clinical data (including image, narrative, and real-time monitoring data), molecular technologies (genomics principal among them), and the availability of devices and wearables to provide high-resolution data streams will dramatically expand the availability of detailed phenotype and environmental data not previously available at this scale. Applications of machine learning approaches could result in new taxonomies of disease through genomic, phenomic, and environmental predictors.

Routine clinical genomics to guide prevention, diagnosis, and therapy

Today, clinical genomic analysis is typically performed only when evaluating certain cancers or when a rare genetic disease is suspected, and many commonly ordered tests only evaluate a few genetic loci. Moving forward, whole-genome approaches will become a routine, early step in the understanding, prevention, detection, and treatment of common and rare diseases.

Rare diseases will increasingly be diagnosed using genomic investigation as a cheaper and more efficient alternative to targeted approaches. Early genome sequencing can solve diagnostic dilemmas and uncover "hidden" Mendelian diseases such as unexplained kidney disease, atypical diabetes, or unexplained development delay (Turro et al., 2020). Some of these Mendelian diseases point to specific new treatments and screening strategies that could dramatically improve health, such as sulfonylureas for young diabetic patients with *HNF1A* mutations or specific causes of liver or kidney failure.

The last decade has also shown that many common conditions, such as diabetes or hypertension, can be associated with genetic risks at thousands of loci, often found using huge genetic studies aggregating data across hundreds of thousands of participants. While many of these genetic loci may have very small

Table 1. Envisioning how precision medicine will affect clinical medicine and research in the next decade

	Where we are today	Where we will be in 2030
Clinical applications		
Genomics for disease	Primarily limited to rare disease and select cancers.	Genomics is routine. Genetic causes and targeted therapies are discovered for many “common” diseases. Microbiome measures are routinely included.
Pharmacogenomics (PGx)	Common in cancer and within select applications of older medications at select sites.	Genome-aware EHRs make PGx easy and automatically update rules from central guidelines. New PGx associations discovered from clinical data.
Genomics for healthy individuals	In research, whole-genome sequencing and search for mutations in one of the ACMG59 genes, present in about 3% of people. Variant interpretation is hard.	ACMG59 grows to > 200, variant interpretation improved by huge, diverse sequenced populations. Cell-free DNA becomes a mainstay of cancer screening
EHRs	Episodic capture from healthcare without robust genomics support. EHR data is essentially not portable.	Genome- and device- enabled. Data can be easily moved between EHRs and to participant apps.
Environmental influences on health	Patient-reported habits and exposures	Geocode-based exposure linkage Real time monitoring of multiple environmental exposures Precision nutrition
Wearable sensors	Ad hoc use of activity monitors	Continuous monitoring of physical activity, sleep, metabolic parameters
Research applications		
Population demographics	>80% European ancestry	>50% non-European ancestry
Routinely available data	Surveys of health conditions, lifestyle, behavior, and diet. GWAS data, lab assays, structured EHR data, and geocoded exposure linkages.	Whole genomes, lab assays, surveys, full EHRs, environmental, genomic and sensor data. Includes imaging, narrative, geocoded, and continuous monitoring approaches to clinical care, activity, precision nutrition, and environment.
Size of cohorts used in analysis	Up to 500K, data downloaded and manually harmonized to sets of several million	>100M using cloud-based federated analyses facilitated by common standards
Largest genomic studies performed on a trait	>1M (GWAS)	>50M (GWAS) >2M (WGS)
Cost of a whole genome	\$500	\$20*

*Sequencing costs have often fallen faster than Moore’s law. Using Moore’s law, sequencing costs would be 1/32 of US \$500, or \$15.63.

genetic effect sizes (with odds ratios < 1.01), they point to pathways involved in disease pathogenesis that may have significant therapeutic implications. Furthermore, weighted aggregations of genetic variants into polygenic risk scores can achieve similar predictive as rare Mendelian disease variants (Khera et al., 2018). Moreover, use of polygenic risk scores may allow providers to risk-stratify individuals who would otherwise be missed by traditional screening approaches, thereby identifying new populations for treatment or screening.

We anticipate that diverse genetic causes and targeted therapies will be uncovered for many common diseases, which could lead to more specific treatment and prevention for the patient and family members. We will likely also discover that many genetic diseases occur on a spectrum of severity, penetrance, and expressivity, guided by the severity of different genetic variants, lifestyle, and environmental interactions. This concept is captured by the scientific agenda of the International Common Disease Alliance. Classic examples include different classes of *CFTR* mutations with cystic fibrosis or *SERPINA1* variants with alpha-1 antitrypsin deficiency, both of which can present with different manifestations and at varying ages given the genetic variant, habits (e.g., smoking), and exposures (e.g., hepatitis virus co-infections).

Routine use of sequencing will produce valuable datasets for secondary research, driving a more comprehensive understanding of disease penetrance, variant pathogenicity, and factors influencing variable expressivity of given genetic variants. It will also produce more patients for whom incidental pathogenic variants are discovered. The American College of Medical Genetics and Genomics has identified 59 genes for which incidental findings should be considered for patient return (i.e., the “ACMG59”) (Kalia et al., 2017). These genes include hereditary cancer syndromes, cardiomyopathies, and potentially fatal arrhythmias, for which actions can be taken to mitigate their risk. Today, about 3% of patients harbor pathogenic variants, the vast majority of which were previously unknown to the patient. As genomic knowledge increases, the number of actionable genes and the fraction of the population affected will significantly increase.

Furthermore, pharmacogenomics can improve drug efficacy, reduce adverse events, and reduce cost. In a 2009 interview, one of the authors of this article (F.S.C.) made the comment, ''if everybody's DNA sequence is already in their medical record and it is simply a click of the mouse to find out all the information you need, then there is going to be a much lower barrier to beginning to incorporate that information into drug prescribing'' (Collins, 2009). Over a decade later, we still have a long way to go. While genomics-guided therapies are becoming the standard of care for some cancers, use of germline pharmacovariants to guide prescribing has been adopted by only a few US medical centers. Implementation has been hindered by a lack of ''genomics-enabled'' electronic health records (EHRs), the complexity of the genetics and recommendations, and a lack of clear evidence. Synthesized evidence and recommendations from the Clinical Pharmacogenomics Implementation Consortium, ubiquity of EHRs supporting complex decision support, and common data standards offer promise to accelerate adoption. Some countries have substantially reduced drug-induced Stevens Johnson Syndrome using genetic testing (White et al., 2018). Even considering just five drug-genomic interactions, nearly everyone has a pharmacovariant that would affect drug prescribing (Van Driest et al., 2014).

EHRs as a source for phenomic and genomic research

The key to any longitudinal cohort is detailed phenotype, exposure, and health outcome assessment. Many site-based and national research cohorts now use EHRs and other health data to provide up to decades of extant disease and treatment information that can be repurposed for research, and we only see this use expanding.

Already EHR-based studies have been instrumental to some of the largest genomic studies of clinically relevant findings, some of which are exceeding 1 million individuals (Vujkovic et al., 2020). By providing a systematic collection of health-related information, EHRs provide phenotypes and data and enable novel study designs often not available in research collections. For example, one study demonstrated participants had an average of more than 190 clinical notes, 14 radiological studies, and more than 700 lab tests over an average of about 8 years of follow up (Robinson et al., 2018). The power to discover specific endophenotypes (e.g., cardiac ejection fraction) or emerging phenotypes (e.g., COVID-19), rare and specific phenotypes (e.g., osteonecrosis of the jaw), or to understand specific manifestations of disease (e.g., bronchiectasis) often requires access to complete EHR data.

EHR data require cleaning and harmonization and can reflect clinical and insurance biases. Unstructured EHR data, such as narrative reports or imaging data, often require advanced methods like natural language processing or machine learning to be useful on a population scale. However, all of these tools are increasingly available and applicable, providing access to data on a scale, depth, and detail not feasible with purely research-collected data.

Clinical EHR data can also be combined with participant-provided research data collections to provide a more complete picture of patient outcomes. Research cohorts such as the UK Biobank and *All of Us* have integrated both data sources.

Further, as clinical sequencing grows, the number of genotypes derived from clinical care will rapidly grow to dwarf those available from research use cases. Many genomic studies may no longer need separate research biospecimen collection to perform large-scale genetic studies. Collection of research biospecimens could then shift toward measuring other biomarkers, cell-free DNA, exposures, and epigenomics.

Higher variety, higher resolution phenomics and environmental exposure data for both clinical and research use

The next decade will see the continued growth of research and clinical uses for different ways to measure clinical phenotypes, exposures, and lifestyle. Data linkages to health claims, national vital statistics, and geospatial resources will become more common as will the use of wearable devices to measure activity, physical measurements, and exposures. Surveys can then be more focused on elements not covered by other methods, thereby decreasing participant burden. Activity monitors that take a number of clinical measurements such as single-lead electrocardiograms and oxygen saturation are becoming inexpensive and can be easily shared with providers. Since the vast majority of a patient's life is spent outside the healthcare system, integration of wearable devices and other patient-provided information would augment the EHR and enable greater telehealth capabilities, experienced first at scale during COVID-19. Moreover, integration of these tools could produce a shift in which most health-related data is derived outside of the healthcare setting.

Despite clear evidence of the impact of nutrition on health, diet is an environmental exposure often ignored in much of clinical practice and many research studies. When it is assessed, it is often through episodic and cumbersome surveys (research) or perfunctory summative questions (in most clinical settings). Replacing dietary assessment with data linkages to grocery stores, digital uploads from restaurants, laboratory and microbiome assessments, or machine learning applied to food imaging would provide more feasible, comprehensive capture of dietary habits. A future of precision nutrition, as a type of ''drug,'' offers a powerful new modality for treating and preventing disease (Rodgers and Collins, 2020).

Privacy, participant trust, and returning value

The utility of precision medicine is dependent on broad participation, and broad participation of large populations requires trust, protection of privacy, and a return of value to the participants. We recognize that science has not always been trustworthy or honored all participants equally. Transparency, authentic engagement with communities, and including participants within research governance can improve trust, create participant advocates, and ensure a more thoughtful, culturally sensitive direction. *All of Us* has involved participants in all levels of governance from the beginning and seeks to return value by giving participants generated research data wherever possible, such as genomics results or upcoming COVID-19 serology results.

Participants also need to trust that their data is secure and private. Highly public data breaches, fear of reidentification, and legal concerns about the availability of certain types of information for factors such as insurability can make this challenging. Clear and honest communication with participants is essential in building trust. Legal protections for the data and technological approaches to ensure secure information systems (such as deidentifying and blurring data, controlling access via blockchain, linking data using privacy-preserving hashed identifiers, and analyzing data using homomorphic encryption) also play a role.

Conclusion

The technologies undergirding precision medicine are already transforming care. Transformative molecular treatments have been developed for rare diseases like cystic fibrosis and spinal muscular atrophy. Genomic investigation led to the development of new drugs for hyperlipidemia. In this time of COVID-19, science has been the answer to an existential medical threat. Yet we are reminded that many of the benefits of medicine's advancement have not always been available to all. Biomedical approaches, computation algorithms, and the availability of high-resolution data will dramatically increase over the next decade. Implementation of a bold plan to collaborate internationally, to engage diverse populations of participants and scientists, to deeply measure our populations, to make clinical and research data broadly available, and to implement this knowledge in clinical practice—in a true learning healthcare system—will allow us to achieve the vision of precision medicine for all populations.

REFERENCES

Collins, F. (2009). Opportunities and challenges for the NIH–an interview with Francis Collins. Interview by Robert Steinbrook. N. Engl. J. Med. *361*, 1321–1323.

Hofstra, B., Kulkarni, V.V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., and McFarland, D.A. (2020). The Diversity-Innovation Paradox in Science. Proc. Natl. Acad. Sci. USA *117*, 9284–9291.

Kalia, S.S., Adelman, K., Bale, S.J., Chung, W.K., Eng, C., Evans, J.P., Herman, G.E., Hufnagel, S.B., Klein, T.E., Korf, B.R., et al. (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet. Med. *19*, 249–255.

Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.

Knepper, T.C., and McLeod, H.L. (2018). When will clinical trials finally reflect diversity? Nature *557*, 157–159.

Manolio, T.A., Goodhand, P., and Ginsburg, G. (2020). The International Hundred Thousand Plus Cohort Consortium: integrating large-scale cohorts to address global scientific challenges. Lancet Digit Health *2*, e567–e568.

Mills, M.C., and Rahal, C. (2020). The GWAS Diversity Monitor tracks diversity by disease in real time. Nat. Genet. *52*, 242–243.

Robinson, J.R., Wei, W.-Q., Roden, D.M., and Denny, J.C. (2018). Defining Phenotypes from Clinical Data to Drive Genomic Research. Annu. Rev. Biomed. Data Sci. *1*, 69–92.

Rodgers, G.P., and Collins, F.S. (2020). Precision Nutrition-the Answer to "What to Eat to Stay Healthy". JAMA *324*, 735–736.

Turro, E., Astle, W.J., Megy, K., Gräf, S., Greene, D., Shamardina, O., Allen, H.L., Sanchis-Juan, A., Frontini, M., Thys, C., et al.; NIHR BioResource for the 100,000 Genomes Project (2020). Whole-genome sequencing of patients with rare diseases in a national health system. Nature *583*, 96–102.

US National Research Council (2011). Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease (Washington, DC: National Academies Press (US)).

Van Driest, S.L., Shi, Y., Bowton, E.A., Schildcrout, J.S., Peterson, J.F., Pulley, J., Denny, J.C., and Roden, D.M. (2014). Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing. Clin. Pharmacol. Ther. *95*, 423–431.

Vujkovic, M., Keaton, J.M., Lynch, J.A., Miller, D.R., Zhou, J., Tcheandjieu, C., Huffman, J.E., Assimes, T.L., Lorenz, K., Zhu, X., et al.; HPAP Consortium; Regeneron Genetics Center; VA Million Veteran Program (2020). Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. Nat. Genet. *52*, 680–691.

White, K.D., Abe, R., Ardern-Jones, M., Beachkofsky, T., Bouchard, C., Carleton, B., Chodosh, J., Cibotti, R., Davis, R., Denny, J.C., et al. (2018). SJS/TEN 2017: building multidisciplinary networks to drive science and translation. J. Allergy Clin. Immunol. Pract. *6*, 38–69.

Wilkins, C.H., Schindler, S.E., and Morris, J.C. (2020). Addressing health disparities among minority populations: why clinical trial recruitment is not enough. JAMA Neurol. *77*, 1063–1064.